Interspeech 2024
1-5 September 2024, Kos, Greece

Unified Framework for Spoken Language Understanding and Summarization in Task-Based Human Dialog processing

Eunice Akani[1,2], *Frederic Bechet*[1,3], *Benoit Favre*[1], *Romain Gemignani*[2]

[1]Aix-Marseille Univ, CNRS, LIS UMR7020, Marseille, France
[2]Enedis, Marseille, France
[3]International Laboratory on Learning Systems - ILLS - IRL CNRS, Montréal, Canada

firstname.lastname@lis-lab.fr, romain.gemignani@enedis.fr

Abstract

Dialogue summarization aims to create a concise and coherent overview of a conversation between two or more people. Recent advances in language models have significantly improved this process, but accurately summarizing dialogues is still challenging due to the need to understand the interactions between speakers to capture the most relevant information. This study focuses on goal-oriented human-human dialogues, incorporating task-related information into the summarization process to produce summaries that are more semantically accurate. It explores multitask approaches that combine summarization with language comprehension tasks and introduces new methods for summary selection and evaluation based on semantic analysis. The study tests these methods on the DECODA corpus, a collection of French spoken dialogues from a call center, showing that integrating models and task-related information improves the accuracy of summaries, even with varying levels of word error rates.

Index Terms: dialog summarization, spoken language understanding, multi-task methods

1. Introduction

Advances in spoken language processing, thanks to low word-error-rate automatic speech transcription and generative language models, have improved the quality of outputs for a range of tasks including speech summarization. Conversation summarization consists in generating a short, abstract version of what was said by speakers within a multiparty spoken interaction. Due to a mismatch between transcript and summary genres, errors in the transcript, disfluencies by participants, and the generative process itself, dialogue summarization is affected by *hallucinations* [1] (*i.e.,* when generative models produce texts that contain nonfactual information when compared to the source).

In the context of industrial applications, this problem might be more prevalent due to the necessity to rely on small language models, for cost reduction and energy savings, as well as for the non-disclosure of customer data. Domain knowledge can help assess faithfulness, and could be integrated in summarization systems in order to generate better summaries.

In this paper, we propose a methodology that involves leveraging data augmentation via LLM to improve the quality of generated summaries. Moreover, we seek to assess and enhance the accuracy of conversation summarization by integrating task-specific elements such as the caller's intent and domain-specific named entities. Our study concentrates on call centers, where the caller's intent is represented as a *call type*. We conduct experiments on the DECODA corpus[2], which stands out as one of the few extensive human-human spoken dialogue datasets collected from authentic call centers. This corpus is valuable due to its annotations for call types and named entities. To our knowledge, there is no comparable large English speech dataset for research featuring goal-oriented human interactions and semantic annotations.

2. Related work

A study conducted by [3] found that 30% of summaries generated by text summarization systems contained incorrect information, known as "hallucinations" [4]. Approaches to assessing the faithfulness of summaries have included textual entailment [5, 4], entities analysis [6, 7], and a verification of answers to questions derived from the summary [8].

Dialogue summarization has become increasingly popular, but models are limited by the structure of dialogues and the diversity of input data, including customer service conversations and technical discussions. To address this, methods using auxiliary information, such as dialogue acts [9] or domain terminology [10], have been proposed. Newer datasets, like SAMSum [11], present new challenges, such as reporting participant behavior in conversations. One limitation of such datasets is that they are not supported by actual speech, and consist in synthetic conversations imagined by annotators.

There are fewer studies on hallucination in dialogue summarization compared to text summarization. [1] found that 35% of SAMSum dataset summaries were inconsistent with the source dialogues. [12] identified eight types of factual errors in dialogue summarization, while [1] identified six types. They used a summary model based on conditional generation probabilities to distinguish between positive and negative summaries and evaluate the model's faithfulness.

Task-oriented dialogue summarization, which involves conversations aimed at accomplishing specific tasks, has emerged as a recent task. To accurately capture the essence of conversations, summaries must reflect participants' goals, procedures, named entities, and other relevant factors. Several corpora have been proposed for the task, including TODSUM [13], and DECODA [2].

3. Data augmentation for spoken conversation summarization

There are two main approaches to generating summaries from audio conversations: *pipeline methods* involve performing automatic transcription followed by automatic text summarization applied to these transcriptions while *end-to-end methods* [14], directly generate summaries from audio inputs without the intermediary step of transcription.

Pipeline methods for training, validating, and testing models necessitate three key resources: audio conversations relevant

10.21437/Interspeech.2024-2276

to the system's target domain, textual transcriptions of these conversations, and examples of text summaries on the transcriptions. The amounts of these required resources can vary significantly based on system specifics like pre-training and adaptation processes.

In contrast, end-to-end methods rely on having large datasets of audio conversation and summary pairs. When such datasets are not readily available, data augmentation techniques, including speech synthesis, may be employed to generate these pairs, as suggested by [15].

This study focuses on pipeline systems for their flexibility in using various resources and models for text or audio processing separately. Pipeline systems offer easy development and analysis without modality alignment.

We describe a realistic scenario where one has access to a corpus consisting solely of raw audio conversations, without any annotations or transcriptions, except for a small subset with target summaries. We show how generic tools for speech transcription and automatic summarization, powered by Large Language Models (LLM), can be leveraged for data augmentation. The methodology involves:

1. Applying a generic speech transcription system across the entire conversation corpus to produce textual transcriptions.
2. Processing these transcriptions with an instruction-based LLM in a few-shot learning mode by feeding the system a few examples of target summaries and prompting it to generate summaries that resemble these targets.
3. Fine-tuning a summary generation system on the dataset including both the automatically generated summaries and the target summaries.

Creating a dedicated summary system is more practical and cost-effective than using a generic LLM. It provides better control over output and allows for more sophisticated strategies, like combining automatic speech understanding and summarization, as discussed in the following section.

4. Integrating Spoken Language Understanding and Summarization

In this section, we describe two methods for integrating semantic task-specific information into the summary generation process. Since we are dealing here with goal-oriented human-human dialogues, we can use the same kind of semantic information that has been proposed for performing Spoken Language Understanding in human-machine dialog systems for tasks such as transport reservation (e.g., ATIS corpus) or for restaurants or hotels (e.g., MEDIA corpus). In this type of study, three semantic levels are generally defined [16]:

- **domain**: the domain represents the semantic context of the dialogue. For instance, in the ATIS corpus, it involves tasks like booking flight tickets. In our study, it pertains to public transportation in Paris for the DECODA corpus.
- **intent**: intent denotes the nature of a request in human-machine communication, such as confirmation or information inquiry. While typically associated with a single utterance, in the DECODA corpus, are assigned to entire dialogue, representing *call-types* like *itinerary request* or *lost item claim*.
- **entity/value pairs**: these pairs represent semantic relations in intents, like a destination in an itinerary request. In the DECODA corpus, entities are locations, bus numbers, time, and service ID numbers, ...

Since we have only one domain in our application corpus, we will only consider the call-types and entities levels. We propose two methods for performing SLU/summarization: during or after the generation process.

4.1. Integrating semantic information during the generation process

Call types and concept labels are not directly present in conversation recordings or transcripts; they need to be inferred. This can be done either through a separate prediction system before generation, integrating them into input data, or during summarization via a multitask approach.
Let D be the input dialogue, S the generated summary, C the call-type, and E the set of entities appearing in the summary. We consider the following methods:

1. **Baseline**: No explicit semantic information is used to generate the summary.
2. **Pipeline**$_C$: Predict call-type C from dialogue D such as $C = intent(D)$, then condition summary generation on C. $S = \text{summary}(D, \text{call-type}(D))$
3. **Multitask**$_{\{C,E,CE\}}$: Generate both semantic labels such as call-type or entities, and summary directly from conversation transcript, resulting in three systems generating the call type, the entities or both prior to generating the summary. $\{C, E, CE\}, S = \text{summary} \circ \text{semantics}(D)$.

In our experiments, a language model was fine-tuned on an automatic summarization task corresponding to each scenario.

4.2. Semantic information as a summary selection process

Alongside incorporating semantic information directly into the summary generation process, we suggest utilizing it post-generation to choose the most semantically reliable summary as per our models. By tweaking parameters like temperature in text generation, multiple outputs with different characteristics can be obtained. We propose a selection method based on call-type prediction and the risk of hallucination on task-related entities.
Call-type prediction: We use a text classifier to predict the call type of a generated summary and compare it to the call type predicted on the entire conversation transcription. We hypothesize that if the predicted call type for the summary closely matches that of the complete transcript, then the summary is likely to be semantically coherent in terms of call types. To handle multiple call types and uncertainty in call type classification, we avoid binary comparison. Instead, we compute a divergence between the probability distribution on all the call-type for the summary and dialogue classifiers. We use the Kullback-Leibler (KL) divergence [17] between these probability distributions of call-types. The KL divergence is a statistical distance measure that quantifies the dissimilarity between two probability distributions. It evaluates the difference between a probability distribution and a reference distribution.
For n the number of call-types, let $G = \{g_1, ..., g_n\}$ be the probability distribution given by the call-type classifier on the generated summary and $R = \{r_1, ..., r_n\}$ the one obtained by the classifier on the entire conversation, the KL divergence between G and R is defined as follows:

$$D_{\text{KL}}(G \parallel R) = \sum_x G(x) \log \left(\frac{G(x)}{R(x)} \right) \quad (1)$$

It is now possible to select the summary that minimizes the D_{KL} distance among the set of generated summaries.

Reducing the entity hallucination risk: We define *hallucination risk* for task-related entities as a selection criterion in previous work [18]. When a summary generation system produces a named entity not in the original document, it increases the risk of a model's over-generation error. Our study proposes the following method to quantify this risk: a NER system is automatically applied to the transcriptions and the generated summaries. We call *NEHR* for *Named Entity Hallucination Risk*, the hallucination risk on named entities [18]. It is calculated as the proportion of entities in the summary that are not present in the conversation. In enhancing the fidelity of the summary, we also use NEHR as a selection criterion alongside D_{KL}.

5. Experiments

5.1. The DECODA corpus

The DECODA corpus [2] contains spoken conversations between several agents of the Paris Transport Authority customer service (RATP) and users of Paris buses and metro lines. To each conversation is associated a manual text transcription and a short summary, called synopsis, which provides a brief overview of the main events of the conversation, including the objectives of the participants and the resolution process. The DECODA corpus covers a variety of call-types, such as *Traffic Information*, *Itinerary*, *Lost/Found Objects*, *Subscription*, *Schedules* and *Tickets*, as documented in [19]. Additionally, the corpus includes entities that belong to a domain ontology, such as *Product, Transport, and Schedule*, among others. DECODA consists of three parts, with annotated synopses available only in parts 1 and 3. Synopses in part 3 are longer, more detailed, and crafted in a literary style, whereas those in part 1 are synthetic and less literary. In this study, as presented in section 3, we decided to keep only a small number of human synopses for system development and use data augmentation for unlabelled data. From the 500 dialogues of the part 3 corpus, we kept 200 dialogues as an evaluation corpus (*test*), 200 for fine-tuning the summarization system (*train-H* for *human* annotated data) and 100 as a validation corpus to adjust the generation parameters. In addition to these gold synopses, we created a corpus of automatic summaries produced by a prompt-based Large Language Model from OpenAI (ChatGPT-3.5), on the transcriptions of the part 1 and part 2 DECODA corpus. We called this corpus *train-A* for *augmented*. Table 1 shows statistics for the training and test sets. Notably, synopses in the test set are generally longer than those in the training set. This disparity arises because the augmented data doesn't consistently mimic the style of human summaries in DECODA, given its 1-shot approach.

Statistics	Train-H	Train-B	Train-H+A	Test
# dialogs	200	697	1390	200
# avg dialog size	545.0	459.28	470.3	495.6
# avg synopsis size	55.3	29.39	47.9	52.7

Table 1: *Decoda data distribution in the train (human and augmented) and the test set. Train-B consists in summary of train-H and original summary from part 1.*

5.2. Summarization and classification models

- **Automatic Summarization**: We trained the summarization systems using BARThez [20], a sequence-to-sequence model pre-trained on various French corpora. It was introduced for the task of automatic text summarization. BARThez is a transformer-based model built on BART architecture [21, 22]. It comes in two versions: base and large. We used the base version with 6 encoder and 6 decoder layers. The pre-trained model provided by the authors and available on the Hugging Face library[1] was used for training the models. We used a learning rate of 5×10^{-5} with AdamW optimizer and set the maximum size for conversations at 1024, and 128 for synopses. Each model was trained over 15 epochs, saving only the one that minimized the loss on the validation set. As automatic transcription lacks speaker role, we've omitted speaker IDs and retained raw text, with speech turns separated by line breaks.

  To train the **Multitask**$_C$ model, we concatenate the call type label to the synopsis for each conversation in the training set. The call types associated with each conversation are required as input for the **Pipeline**$_C$ model. A call type classifier, trained through k-fold cross-validation, predicts call types on 25% of the data not initially used for training. This process is repeated four times to obtain predictions for the entire training set, which are then used to train the summarization system. A marker separated the call types and conversations to help the model consider call types when predicting summaries. All our summarization experiments are evaluated with the ROUGE [23] and BERTScore [24] metrics.
- **Calltype Classification** Two classifiers based on CamemBERT-base [25] were trained to classify the call-type: one taking as input automatic conversation transcripts [Conv. $\rightarrow$ call-type] and the other one, generated synopses [Syn. $\rightarrow$ call-type]. Each model was trained over 15 epochs and the model with the minimal loss on the validation data was retained.
- **Domain-Specific Named Entity Recognition** DECODA has 14 domain-specific named entities, among which are phone number, price, product type, transport type, etc. We trained CamemBERT-base [25] for the NER task and obtained a micro F1 and a macro F1 of 0.93 and 0.84, respectively, using the Seqeval [26]. The detected entities are used in the evaluation metrics.

5.3. Evaluation Results

Impact of data augmentation
To validate our data augmentation strategy, we computed Rouge-L and BERTScore obtained on the manual transcriptions of our test partition by 3 different models: It can be noticed that Barthez trained on Orange-sum, a text summarization dataset, can't give better result (see table 2). This means that the automatic summarization task doesn't seem to be transferable to dialogue summarization without prior training.
Call type Classification
After training the two call type classifiers, the results are recorded in Table 3. We can see that the classifiers have similar results. We also recorded in table 3 the performance of the multitasks Multitask$_{C,CE}$. Multitask$_C$ performed better than Multitask$_{CE}$. This can be due to the huge amount of information to generate.
Impact of Word Error Rate on summarization performance
In an industrial context, as manual transcriptions are difficult to acquire without annotation, the use of ASR systems is a good alternative. We evaluated different sizes of WhisperX [27], an ASR system, and computed its WER score based on manual

[1] https://huggingface.co/moussaKam/barthez

System	train data	Rouge-L	BERTScore
Barthez	OrangeSum [20]	14.36	7.82
Barthez	train-H	23.59	33.12
Barthez	train-H+A	29.11	38.90
chatGPT3.5	-	28.93	37.30

Table 2: *RougeL and BERTScore on the test partition on the manual transcription of DECODA with Barthez fined tuned on the different datasets (OrangeSum, train-H et train-A). Performance of chatGPT3.5 is also presented.*

Systeme	Acc.	W-F1
Conv. $\rightarrow$ call-type	81	80
Syn. $\rightarrow$ call-type	80	80
Multitask$_C$	77	75
Multitask$_{CE}$	73	71

Table 3: *Accuracy and Weighted F1 Score of call types classifiers and generation systems*

transcriptions. We trained a dialogue summarization model using these transcriptions to observe the impact of the WER score on the summarization score. The results are consigned in Table 4. Regardless of the system used, the performance of ASR systems does not match that of a summarization system trained on manual transcriptions. We observed that as the WER score decreases, the ROUGE and the BERT scores also improve. Our findings suggest that the WER score has an impact on dialogue summarization. We will use the automatic transcription generated by Whisper large for upcoming experiments as it is the closest to manual transcription.

Transcript	WER	Rouge-L	BERTScore
Manual	0.00	29.11	38.90
Whisper tiny	76.37	24.53	34.38
Whisper small	46.37	26.99	36.95
Whisper large	40.02	27.67	37.14

Table 4: *WER score on automatic transcription from whisper; Rouge-L and BERT-scores on summaries generated by Barthez finetuned and evaluated on automatic transcriptions*

Impact of semantic information during training

We trained pipeline and multitask models (section 4.1) on the train-H+A dataset with WhisperX large transcription and reported results on testing dataset into table 5. In addition to automatic summarization scores, we computed the acc-ref score representing the accuracy between reference call types and those predicted by [Syn. $\rightarrow$ call-type] for each generated summary. We also provided entity precision comparisons between the reference and generated summaries. Pipeline$_C$ improved acc-ref on call type while having similar results in terms of ROUGE-L and ENT-prec. Contrary to our expectations, Multitask$_{CE}$ does not improve the ENT-prec. This can be due to the complexity of the task, as it had a low accuracy score in generated call types. Multitask$_C$ and Multitask$_E$ yielded comparable results to the baseline. This could possibly mean that integrating semantic information such as call types and named entities for multitask generation may not significantly enhance the semantic quality of the generated summary.

System	RL	BS	acc-ref	ENT-prec
Baseline	27.67	37.14	0.76	0.53
Pipeline$_C$	27.61	36.97	0.79	0.54
Multitask$_C$	27.41	37.16	0.76	0.51
Multitask$_E$	28.03	37.21	0.75	0.46
Multitask$_{CE}$	27.51	37.02	0.74	0.51

Table 5: *Summaries evaluation Rouge-L, BERT-scores. acc-ref refers to call type prediction from synopsis (ref=oracle synopsis) and ENT-prec the precision of entities between generated summaries and synopses.*

Impact of semantic information as a summary selection

As Pipeline$_C$ increased the acc-ref, we used the model trained on it to generate various summaries using sampling decoding strategies and selecting the one according to the selection metrics presented in section 4.2. We report automatic results in table 6 and denote as **Pipeline$_C$ - kl** for summaries selected using D_{KL} criterion and **Pipeline$_C$ - nehr** for summaries selected using NEHR criterion. We also generated various summaries using ChatGPT3.5 to see the performance of LLM. For Pipeline$_C$, we have similar automatic text summarization score while acc-ref and ENT-prec vary. We can see that D_{KL}-based selection increases acc-ref but reduces entity precision, while NEHR-based selection increases both values. For ChatGPT3.5 it is D_{KL}-based selection that gives the best result in terms of acc-ref and entity precision. This suggests that the semantic impact of selection criteria could vary between a small model and an LLM, but further experiments are needed to confirm this.

System	RL	BS	acc-ref	ENT-prec
Pipeline$_C$	27.61	36.97	0.79	0.54
Pipeline$_C$ - kl	27.77	36.84	0.82	0.52
Pipeline$_C$ - nehr	27.65	36.36	0.81	0.58
ChatGPT3.5	26.62	35.41	0.79	0.63
ChatGPT3.5 - kl	26.11	34.35	0.83	0.65
ChatGPT3.5 - nehr	26.37	34.80	0.80	0.60

Table 6: *Summaries evaluation of summary selection method. ChatGPT3.5 results are included*

6. Discussion and Conclusion

We examined the impact of different elements including data augmentation, automatic transcription, semantic usage on the conversation summary generation. We saw that using LLM to generate a part of training synopses increases the performance of model according to ROUGE score and BERTScore. If manual transcription is not feasible, automatic transcriptions with a WER below 40% could be a viable alternative to bridge the gap in annotated data as the ROUGE score is relatively high. Combining call types and conversation as input appears to enhance conversation semantics, as indicated by improved accuracy between reference and predicted call types from the generated summary. Additionally, using NEHR in this model boosts named entity precision. However, manual evaluation is required to confirm the effectiveness of this selection criterion in terms of fidelity and informativeness. As not all corpus metadata has been utilized, we aim to incorporate other aspects, such as conversation structure, to further enhance the fidelity of generated summaries.

7. References

[1] B. Wang, C. Zhang, Y. Zhang, Y. Chen, and H. Li, "Analyzing and evaluating faithfulness in dialogue summarization," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: ACL, Dec. 2022, pp. 4897–4908. [Online]. Available: https://aclanthology.org/2022.emnlp-main.325

[2] F. Bechet, B. Maza, N. Bigouroux, T. Bazillon, M. El-Bèze, R. De Mori, and E. Arbillot, "DECODA: a call-centre human-human spoken conversation corpus," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 1343–1347. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2012/pdf/684_Paper.pdf

[3] Z. Cao, F. Wei, W. Li, and S. Li, "Faithful to the original: Fact aware neural abstractive summarization," *ArXiv*, vol. abs/1711.04434, 2018.

[4] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in *Proceedings of the 58th Annual Meeting of the ACL*. Online: ACL, Jul. 2020, pp. 1906–1919. [Online]. Available: https://aclanthology.org/2020.acl-main.173

[5] T. Falke, L. F. R. Ribeiro, P. A. Utama, I. Dagan, and I. Gurevych, "Ranking generated summaries by correctness: An interesting but challenging application for natural language inference," in *Proceedings of the 57th Annual Meeting of the ACL*. Florence, Italy: ACL, Jul. 2019, pp. 2214–2220. [Online]. Available: https://aclanthology.org/P19-1213

[6] F. Nan, R. Nallapati, Z. Wang, C. Nogueira dos Santos, H. Zhu, D. Zhang, K. McKeown, and B. Xiang, "Entity-level factual consistency of abstractive text summarization," in *Proceedings of the 16th Conference of the European Chapter of the ACL: Main Volume*. Online: ACL, Apr. 2021, pp. 2727–2733. [Online]. Available: https://aclanthology.org/2021.eacl-main.235

[7] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," vol. 55, no. 12, mar 2023. [Online]. Available: https://doi.org/10.1145/3571730

[8] E. Durmus, H. He, and M. Diab, "FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization," in *Proceedings of the 58th Annual Meeting of the ACL*. Online: ACL, Jul. 2020, pp. 5055–5070. [Online]. Available: https://aclanthology.org/2020.acl-main.454

[9] C.-W. Goo and Y.-N. Chen, "Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts," 2018.

[10] J. J. Koay, A. Roustai, X. Dai, D. Burns, A. Kerrigan, and F. Liu, "How domain terminology affects meeting summarization performance," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 5689–5695. [Online]. Available: https://aclanthology.org/2020.coling-main.499

[11] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, "SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization," in *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Hong Kong, China: ACL, Nov. 2019, pp. 70–79. [Online]. Available: https://aclanthology.org/D19-5409

[12] X. Tang, A. Nair, B. Wang, B. Wang, J. Desai, A. Wade, H. Li, A. Celikyilmaz, Y. Mehdad, and D. Radev, "CONFIT: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning," in *Proceedings of the 2022 Conference of the North American Chapter of the ACL: Human Language Technologies*. ACL, 2022. [Online]. Available: https://doi.org/10.18653%2Fv1%2F2022.naacl-main.415

[13] L. Zhao, F. Zheng, K. He, W. Zeng, Y. Lei, H. Jiang, W. Wu, W. Xu, J. Guo, and F. Meng, "Todsum: Task-oriented dialogue summarization with state tracking," 2021.

[14] R. Sharma, S. Palaskar, A. W. Black, and F. Metze, "End-to-end speech summarization using restricted self-attention," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8072–8076.

[15] K. Matsuura, T. Ashihara, T. Moriya, T. Tanaka, A. Ogawa, M. Delcroix, and R. Masumura, "Leveraging large text corpora for end-to-end speech summarization," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[16] J. Lee, D. Kim, R. Sarikaya, and Y.-B. Kim, "Coupled representation learning for domains, intents and slots in spoken language understanding," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 714–719.

[17] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951, publisher: Institute of Mathematical Statistics. [Online]. Available: https://www.jstor.org/stable/2236703

[18] E. Akani, B. Favre, F. Bechet, and R. Gemignani, "Reducing named entity hallucination risk to ensure faithful summary generation," in *Proceedings of the 16th International Natural Language Generation Conference*, C. M. Keet, H.-Y. Lee, and S. Zarrieß, Eds. Prague, Czechia: Association for Computational Linguistics, Sep. 2023, pp. 437–442. [Online]. Available: https://aclanthology.org/2023.inlg-main.33

[19] J. Trione, "Extraction methods for automatic summarization of spoken conversations from call centers (méthodes par extraction pour le résumé automatique de conversations parlées provenant de centres d'appels) [in French]," in *Proceedings of TALN 2014 (Volume 4: RECITAL - Student Research Workshop)*. Marseille, France: Association pour le Traitement Automatique des Langues, Jul. 2014, pp. 104–111. [Online]. Available: https://aclanthology.org/F14-4010

[20] M. Kamal Eddine, A. Tixier, and M. Vazirgiannis, "BARThez: a skilled pretrained French sequence-to-sequence model," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: ACL, Nov. 2021, pp. 9369–9390. [Online]. Available: https://aclanthology.org/2021.emnlp-main.740

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[22] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the ACL*. Online: ACL, Jul. 2020, pp. 7871–7880. [Online]. Available: https://aclanthology.org/2020.acl-main.703

[23] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: ACL, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[24] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[25] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot, "CamemBERT: a tasty French language model," in *Proceedings of the 58th Annual Meeting of the ACL*. Online: ACL, Jul. 2020, pp. 7203–7219. [Online]. Available: https://aclanthology.org/2020.acl-main.645

[26] H. Nakayama, "seqeval: A python framework for sequence labeling evaluation," 2018, software available from https://github.com/chakki-works/seqeval. [Online]. Available: https://github.com/chakki-works/seqeval

[27] M. Bain, J. Huh, T. Han, and A. Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," *INTERSPEECH 2023*, 2023.